IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834, p- ISSN: 2278-8735.Volume 19, Issue 5, Ser. 1 (Sept. – Oct. 2024), PP 01-07
www.iosrjournals.org

Application Of Genomic Signal Processing To Detect Cancer Cells

Syeda Maryam Fathima[1], Amreen Sulatan[2], Nuzath Sultana[3],
Rabiya Mohammadi Syeda[4]
[1,2,3,4]*(B.E Electronics And Communication Undergraduate, Muffakham Jah College Of Engineering And Technology, Hyderabad, Telangana, India)*

Abstract:
In the field of signal processing, a new area of research has been introduced namely genomic signal processing (GSP). GSP processes genes, proteins, and DNA sequences using various hidden signals. As some genetic abnormalities turn into cancer diseases, proper understanding, and analysis of genes and proteins may lead to a new horizon in cancer genomic study. In genomic signal processing, identifying and classifying the diseased gene is a great challenge to researchers. Hence in the present paper, the crucial job of gene identification and classification is attempted for cancer detection. Our project is implemented in MATLAB R2019a using the bioinformatics toolbox. Where the DNA sequences obtained from the NCBI are processed and numerically mapped before extracting the exons using the period-3 property which is done using an anti-notch filter and STFT. Digital filters are used for noise reduction and increased accuracy.

Keywords: *Genomic signal processing, Cancer detection, DNA Sequence, NCBI*

Date of Submission: 26-08-2024 Date of Acceptance: 06-09-2024

I. Introduction

Over the past decade, significant discoveries have been made that provide a better understanding of genetic basis of cancer disease. It has been understood that DNA plays an important role in the study of cancer. DNA is a molecule that contains the instructions an organism needs to develop, live and reproduce. These instructions are found inside every cell and are passed down from parents to their children. DNA is composed of smaller components called *nucleotides.* There are four types of nucleotides denoted by the letters Adenine (A), Thymine (T), Guanine (G), and Cytosine(C). Regions in a genome that code for proteins are called genes. Genes are further split into coding regions called *exons* and noncoding regions called *introns.* Accurate location of exons in genomes is very important for understanding life processes. Hence this project aims at identification of cancer genes through genomic signal processing.

The most fascinating thing about life is not its diversity but its fundamental building block. All living organisms are made up of microscopic fundamental biological structures called cells. Even though the cells are very tiny, each of them is, in turn, made up of complex cellular substructures. Each living cell generates its energy and synthesizes its own macromolecules required for other biological processes. Some organisms such as bacteria and baker's yeast are unicellular, i.e., they contain only a single cell. Most other organisms are multicellular, containing many different types of cells. Living cells may be divided into two types, the simpler prokaryotic cell, and the more complex eukaryotic cell. By definition, prokaryotes are those organisms whose cells are not subdivided by membranes into a separate nucleus and cytoplasm. All prokaryote cell components are located together in the same compartment.

All living cells contain the essential chemical and structural components necessary for supporting life. For example, each bacterial cell has a single chromosome carrying a full set of genes providing it with the genetic information necessary to operate like a living organism. More complex organisms have genetic information much more than that of bacteria. Humans have two duplicate sets of 23 different chromosomes making a total of 46 chromosomes. Each chromosome is made up of a macromolecule, Deoxyribonucleic acid (DNA). DNA is a nucleic acid that contains the hereditary information in living organisms. A DNA molecule is composed of two complementary strands coiled around each other forming a double helix structure, as shown in Fig. 1.1.

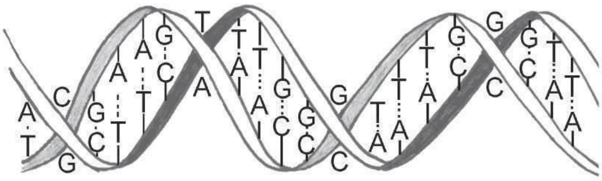


fig 1.1: schematic of DNA double helix.

Each strand consists of a long chain of nucleotides, which are identified by the four nitrogenous bases: Adenine(A), Cytosine(C), Guanine (G), and Thymine (T). Accordingly, a DNA strand is represented by a character sequence consisting of the four alphabet letters {A, C, G, and T}. DNA sequences are divided into genes and intergenic regions. Genes in eukaryotes are further segmented into protein-coding regions (exons) and non-coding regions (introns), as illustrated in Fig. 1.2.

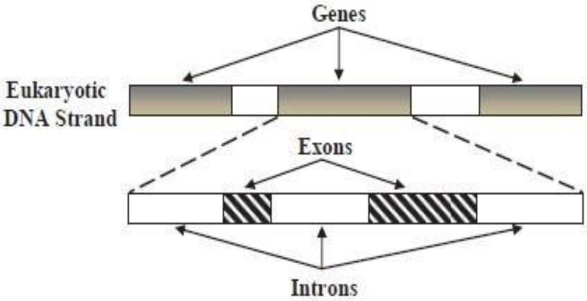


fig 1.2: schematic of eukaryotic DNA structure.

Ribonucleic acid (RNA) is a working copy of DNA resulting from a process known as transcription based on the information contained in DNA. RNA is very similar to DNA except that in RNA the nucleotide uracil (U) replaces thymine (T) in DNA, and RNA is normally found as a single-stranded molecule, whereas DNA is double-stranded. From the viewpoint of genetic information, T in DNA and U in RNA are equivalent. The main job of RNA is to transfer the genetic information contained in DNA from the nucleus to the ribosome for the creation of proteins. This process prevents the DNA from having to leave the nucleus. This process keeps the DNA and genetic code protected from being corrupted

II. Methods and Materials

Source of Input

The National Center for Biotechnology Information (NCBI), established in 1988, is a division of the U.S. National Library of Medicine. Based in Bethesda, Maryland, NCBI maintains crucial bioinformatics resources, including GenBank for DNA sequences and PubMed for biomedical literature. These comprehensive databases are freely accessible online through the Entrez search engine, serving as essential tools for researchers worldwide in biotechnology and biomedicine. For our project, we use these NCBI databases as the primary source of genomic data, ensuring we work with high-quality, standardized information for our analyses.

Technical Approach

The Our project utilizes MATLAB R2016a with its Bioinformatics Toolbox for data processing and analysis. The workflow begins with obtaining DNA sequences from the NCBI database in FASTA format. These sequences are then converted into numerical representations to enable digital signal processing (DSP) techniques, particularly for period-3 detection.

To prepare the DNA sequences for DSP analysis, we employ the Pseudo-EIIP (Electron-Ion Interaction Potential) mapping method. This approach assigns specific numerical values to each nucleotide:
* Adenine (A): 0.1994
* Thymine (T): 0.1933
* Guanine (G): 0.0123
* Cytosine (C): 0.0692

This This numerical conversion is crucial as it transforms the symbolic DNA sequence into a format suitable for advanced signal processing techniques, allowing us to extract meaningful genomic features and patterns.

III. Genomic Signal Processing

Period 3 property

GSP can be used as an effective tool for the analysis of genomic data. The discrete nature of the DNA information, being discrete in both “time” and “amplitude,” invites investigation by digital signal processing (DSP) techniques. Several approaches have been suggested for differentiating between the protein-coding and non- protein-coding regions of DNA. A selection of these approaches is based on genomic signal processing (GSP) techniques. These GSP techniques rely on the phenomenon that protein-coding regions have a prominent power spectrum peak at frequency f = 1/ 3 arising from the length of codons (three nucleic acids). It crucial to identify the segments within the DNA sequence that involved in protein synthesis and is called a coding region of the gene (i.e., exons). The methods are generally used to identify the segment that relies on the period-3 property of genes. 3-base periodicity in a DNA sequence is partly caused by the unbalanced nucleotide distributions in the three coding positions in the sequence. The reason for the unbalanced distribution is that proteins prefer special amino acid compositions and thus nucleotide usage in a coding region is highly biased. The power spectrum of DNA segments corresponding to exons exhibits a strong component at frequency 2π/3 shown in Fig 2.1, known as the period-3 frequency, whereas segments corresponding to introns do not. Exons can, thus, be located by tracking the strength of the period-3 frequency component along the length of a DNA sequence. However, maximizing exons prediction accuracy is still a challenging research issue in DNA sequence analysis.

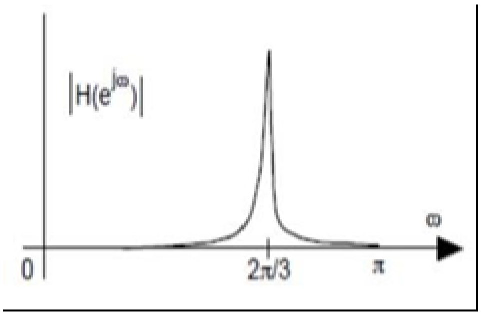


fig 3.1: magnitude response of anti-notch filter

Identifying exon locations in a DNA sequence using DSP techniques is mainly a three-stage process, as illustrated in Fig 3.1.

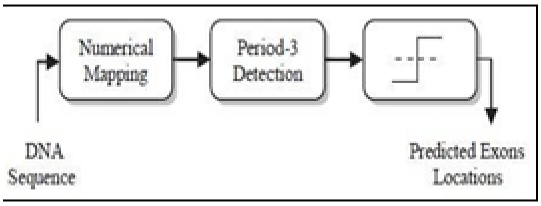


fig 3.2: block diagram of gsp-based exon prediction approach

GSP can be used as an effective tool for the analysis of genomic data. The discrete nature of the DNA information, being discrete in both “time” and “amplitude,” invites investigation by digital signal processing (DSP) techniques. Several approaches have been suggested for differentiating between the protein-coding and non- protein-coding regions of DNA. A selection of these approaches is based on genomic signal processing (GSP) techniques. These GSP techniques rely on the phenomenon that protein-coding regions have a prominent power spectrum peak at frequency f = 1/ 3 arising from the length of codons (three nucleic acids).

To achieve our research objectives, the project involves several key steps. DNA sequences from normal and cancer samples are extracted from databases like HMR195, NCBI, PubMed, Uniprot, and BG570. The process begins by converting symbolic DNA sequences into numerical sequences using the EIIP mapping technique. This numerical data is then analyzed with signal processing methods to identify protein-coding regions. Next, the Discrete Wavelet Transform (DWT) is applied to capture wavelet features, exploiting the 3-base periodicity for detailed approximation. Finally, the sequences are classified as cancerous or normal based on a threshold value. The overall procedure is outlined in the block diagram shown in Fig. 3.3.

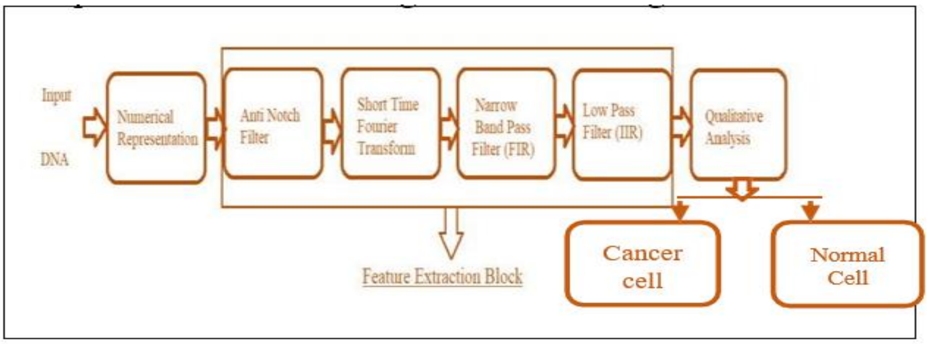


fig 3.3: block diagram

EIIP-based Mapping Technique

Before computational methods can be applied, it isnecessary to convert the A, T, G and C character sequences into numeric sequences. There are many methods for numerical mapping of DNA sequences like fixed mapping method, Physico Chemical Property based Mapping, Statistical Property based Mapping etc. [7, 14, 15]. In this work, we have assigned EIIP value to DNA string to convertinto numerical sequence. The EIIP value for DNA nucleotides is shown in Table 1. EIIP is an energy of delocalized electrons of nucleotides hence this scheme provides more biological information and also reduces computational overhead by 75% [20]. For e.g. if x[n] = [A G C A A G T C A], then on substituting values from Table 1. X[n] = [0.1260 0.0806 0.1340 0.1260 0.1260 0.0806 0.1335 0.1340 0.1260].

Identification of protein-coding regions using anti-notch filters

A major area of research in genomic sequence analysis is the identification of protein-coding regions using the period-3 property. Previously anti-notch filter has been used for this purpose. paper, three anti-notch filters, namely conjugate suppression anti-notch filter, anti-notch filter followed by moving average filter, and harmonic suppression antI-notch filter are proposed to improve the identification accuracy. Conjugate suppression anti-notch filter suppresses the conjugate frequency component, anti-notch filter followed by moving average filter reduces the background noise and harmonic suppression anti-notch filter suppresses the harmonic frequency component. Various DNA-to-numerical mapping ways are compared using the GENSCAN test set, leading to a recommended mapping fashion for detailed analysis with different datasets. The computational complexity of these anti-notch pollutants is assessed and set up to be significantly lower than that of the ST-DFT system. When compared to being anti-notch pollutants at the nucleotide position using standard datasets, the proposed styles demonstrate superior performance, performing in more accurate identification of protein-rendering regions.

Equation 1:

$$y(n) = \sum_{i} |y_i(n)|^2 \quad F=\{A,T,C,G\}.$$

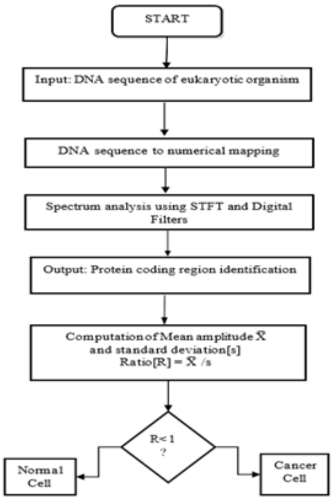

START
Input: DNA sequence of eukaryotic organism
DNA sequence to numerical mapping
Spectrum analysis using STFT and Digital Filters
Output: Protein coding region identification
Computation of Mean amplitude X̄ and standard deviation[s] Ratio[R] = X̄ /s
R< 1 ?
Normal Cell
Cancer Cell

Fig 3.4 Flow Chart Of Cell Identification

A plot of Y (n) will have peaks in the coding regions whereas such peaks are absent in noncoding regions. So, this feature can be utilized for the identification of protein-coding regions in a DNA segment. This anti-notch filter has two poles at Re±jθ and two zeros at ±1. The value of pole radius R should be less than one for stability. The magnitude response and pole-zero plot of the anti-notch filter is shown in Fig. 1. It is observed that the filter has two passbands with center frequencies at 2π/3 and 4π/3 because of the two poles at those frequencies. This second-order IIR anti-notch filter is stable and real.

To classify different groups of cancer genes. The distribution of hydrophilic amino acids in genes is used here as a key feature for their identification. Mutual information of cancer genes based on minimum entropy mapping is estimated for cancer classification.

Once the cancer genes are identified, healthy Homo sapience genes are discarded. The proposed algorithm for classification is applied only to different types of cancer genes. Here cancerous genes are classified by measuring mutual information. The mutual information measures the amount of information of one random variable that contains information about another random variable. It reduces the uncertainty of one random variable due to the knowledge of the other. Mutual information is a parameter between two genes, measuring their degree of dependence. Higher mutual information between two genes implies the existence biological relationship between them. The SNR is one of the figures of merit that describes the quality of a particular analysis technique. The higher the ratio the easier it is to extract information and the more reliable the results. In this work ratio of mean amplitude and standard deviation of signal is considered as figure of merit. The mean amplitude and standard deviation of signal and calculated ratio for all the downloaded sequences with their reference number. It has been shown that the ratio of mean amplitude to standard deviation, less than 1.0 indicates normal cells and more than 1.0 indicates cancer cells.

IV. Results

The study investigates the early detection of cancer by analyzing DNA sequences, particularly focusing on the identification of protein-coding regions. It was observed that protein-coding regions exhibit a distinct "periodicity" or "period-3" behavior, where peaks occur at intervals of k=N/3, a characteristic not found in non-protein coding regions. This property is crucial for identifying exon regions in DNA sequences.

To detect this period-3 behavior, an anti-notch filter is employed, which suppresses the conjugate frequency components. The research analyzed existing techniques and found that by assigning appropriate numerical values to the characters in DNA strings, Digital Signal Processing (DSP) techniques can effectively distinguish between protein-coding (exon) and non-protein coding (intron) regions.

The study further differentiates between normal and cancerous cells by analyzing the power spectrum plot, where spikes are present in cancer cells but absent in normal cells. Parameters such as mean amplitude, standard deviation, and the coefficient of variation are computed to predict the presence of cancer. It was observed that the ratio of mean amplitude to mean frequency is less than 1.0 in cancer cells and greater than 1.0 in normal cells.

The algorithm developed in this study was successfully tested on various DNA sequences from both normal and cancer cells, using data from the NCBI website, confirming its effectiveness in early cancer detection.

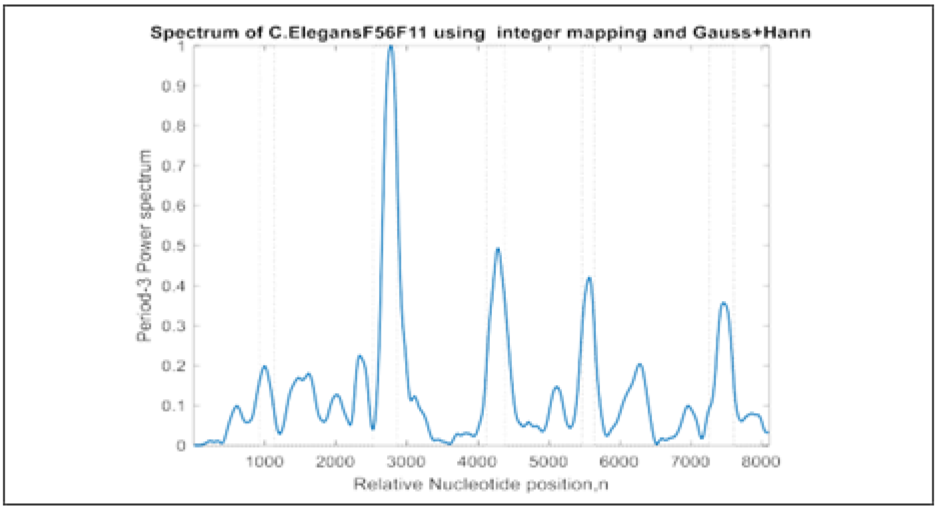


fig 4.1: exon prediction region

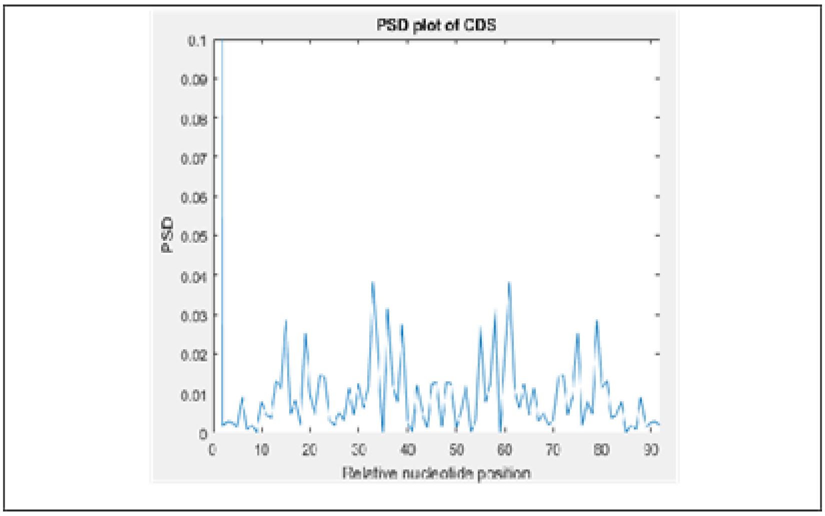


fig 4.2 normal cell CDS plot

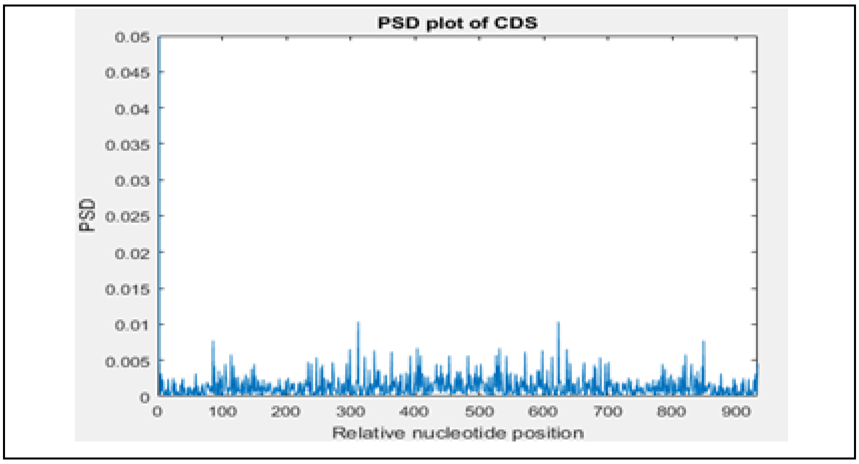


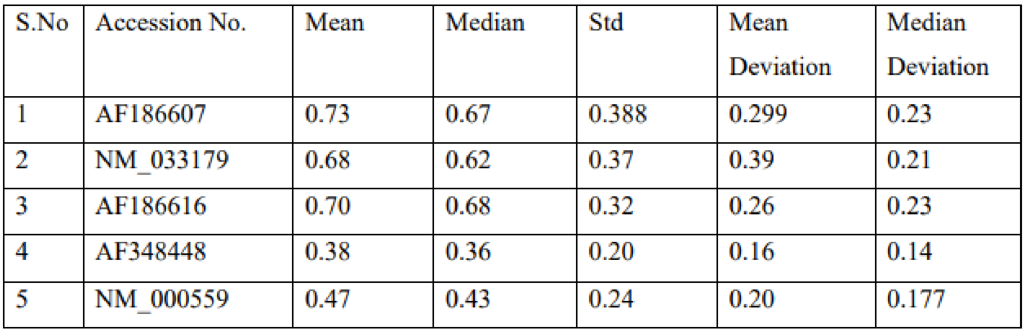
fig 4.3 cancer cell CDS plot

Table no 1 presents statistical parameters with values that are consistently higher in cancerous cells compared to normal cells. These parameters were derived from the results obtained using MATLAB

S.No	Accession No.	Mean	Median	Std	Mean Deviation	Median Deviation
1	AF008216	1.33	1.25	0.70	0.56	0.49
2	NM_004448	1.37	1.27	0.729	0.57	0.488
3	NM_007297	1.55	1.47	0.81	0.63	0.52
4	NM_007299	1.13	1.07	0.59	0.48	0.41
5	NM_001144917	1.24	1.18	0.65	0.52	0.46

Table no 1 Statistical Parameters for cancerous cells

Table no 2 presents statistical parameters for various non-cancerous cells, showing that all values are consistently lower than those observed in cancerous cells. These parameters were derived from computational results obtained using MATLAB.



S.No	Accession No.	Mean	Median	Std	Mean Deviation	Median Deviation
1	AF186607	0.73	0.67	0.388	0.299	0.23
2	NM_033179	0.68	0.62	0.37	0.39	0.21
3	AF186616	0.70	0.68	0.32	0.26	0.23
4	AF348448	0.38	0.36	0.20	0.16	0.14
5	NM_000559	0.47	0.43	0.24	0.20	0.177

Table no 2 Statistical Parameters for non-cancerous cells

V. Conclusion

The prediction is done by computing the parameters such as meanamplitude, standard deviation, and the coefficient of variation. It has been observed that the ratio of mean amplitude to mean frequency is less than 1.0 for cancer cells and more than 1.0 for normal cells. The algorithm is successfully tested for several DNA sequences for both normal and cancer cells with various accession numbers collected from the NCBI website The proposed method efficiently extracts protein-coding regions, or exons, from DNA sequences, which is crucial for early cancer detection. Protein-coding regions exhibit a distinctive "periodicity property" with peaks at k=N/3, a behavior not seen in non-coding regions. This period-3 pattern is pivotal for identifying exons. The method employs an Anti-notch filter to detect this periodicity by suppressing conjugate frequency components.

By assigning numerical values to DNA characters, digital signal processing (DSP) techniques can differentiate between protein-coding (exon) and non-coding (intron) regions. Discrimination between normal and cancerous cells is achieved by analyzing spikes in the power spectrum plot, which are present in cancer cells but absent in normal cells. Parameters such as mean amplitude, standard deviation, and coefficient of variation are computed for prediction. Notably, the mean amplitude-to-mean frequency ratio is less than 1.0 for cancer cells and greater than 1.0 for normal cells. The algorithm has been successfully tested on various DNA sequences from both normal and cancer cells, using data collected from NCBI.

Future advancements could enhance prediction accuracy through advanced filtering, mapping, and transformation techniques. The next step is to apply these methods across various oncogene databases to further refine and generalize the analysis

Acknowledgments

Syeda Maryam Fatima, Amreen Sulana, Nuzath Sultana, and Rabiya Mohammadi Syeda, would like to offer a heartfelt appreciation to our supervisor Mrs. Kamala Kumari, for providing her valuable guidance and for giving us the chance to do this project and study. Lastly, we also would like to thank everyone who helped us complete this project.